Чекпоинт 1: Сбор и подготовка данных

RAG Чат-бот по роману «Мастер и Маргарита»

Команда bookworm

Горбунов Дмитрий Павлович
Ковалева Дарина Евгеньевна
Тельнов Федор Николаевич
Мацаков Борис Вячеславович

Репозиторий: https://github.com/bookworm-itmo/llm-intro-project

Источник данных

Роман М.А. Булгакова «Мастер и Маргарита» в формате FB2, скачан из электронной библиотеки Флибуста. Книга подходит для RAG: большой объём текста, много персонажей и событий.

Сбор данных

Разработан парсер FB2, который:

- Извлекает текст из XML-структуры файла
- Разбивает на главы по паттерну «Глава N»
- Сохраняет текст с привязкой к номерам глав

Исходный файл: `data/master_and_margarita.fb2`

Подготовка для RAG

Chunking — разбиение текста на фрагменты с помощью `RecursiveCharacterTextSplitter`:

- Размер чанка: 800 символов, перекрытие: 100 символов
- Каждому чанку присвоен номер главы
- Результат: **1182 чанка** из 32 глав

Эмбеддинги — векторизация через GigaChat Embeddings (Сбер):

- Размерность: 1024, нормализация L2

Индексация — FAISS IndexFlatIP для косинусного поиска.

Структура данных

Все данные в папке data/:

```
data/
 master_and_margarita.fb2   # Исходная книга
 chunks.parquet             # Чанки (chunk_id, chapter, text)
 embeddings.parquet         # Векторы (chunk_id, embedding[1024])
 chunks_sample.json         # Сэмпл данных
 faiss_index/index.faiss    # Векторный индекс
```

Объём данных

Исходный текст:	~580 000 символов
Глав:	32 + эпилог
Чанков:	1 182
Средний размер чанка:	~705 символов

Сэмпл данных

Доступен по ссылке: data/chunks_sample.json

Пример чанка (глава 1):

> «Никогда не разговаривайте с неизвестными. В час жаркого весеннего заката на Патриарших прудах появилось двое граждан. Первый из них — приблизительно сорокалетний, одетый в серенькую летнюю пару, — был маленького роста, темноволос, упитан, лыс...»

Архитектура

Система состоит из трёх сервисов:

- **Data Service** — парсинг FB2, chunking, подготовка данных
- **RAG Service** — поиск релевантных фрагментов через FAISS
- **LLM Service** — генерация ответов через Claude API

Пользователь взаимодействует с системой через веб-интерфейс (Streamlit).

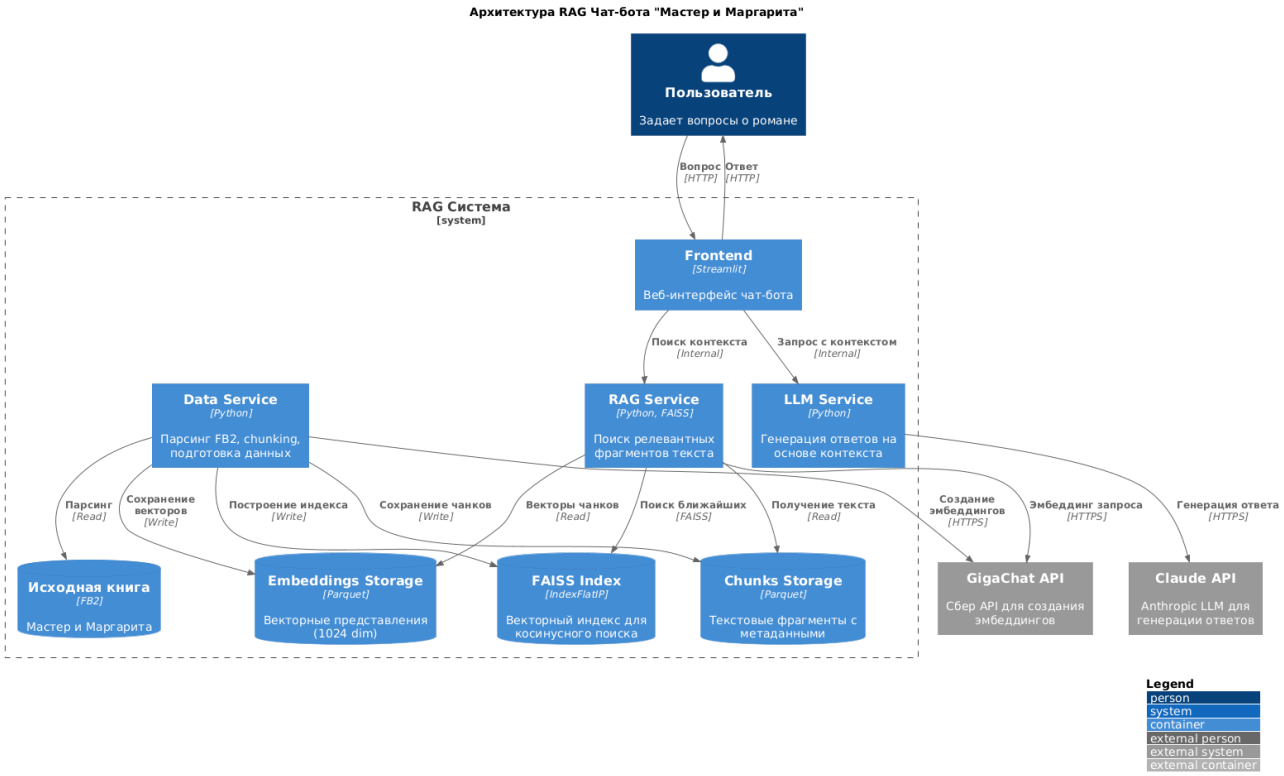

Рис. 1: Архитектура RAG чат-бота